A National Virtual Specimen Database for Early Cancer Detection

Heather Kincaid
Fred Hutchinson Cancer Research Center
hkincaid@fhcrc.org

Daniel Crichton
NASA Jet Propulsion Laboratory
California Institute of Technology
dan.crichton@jpl.nasa.gov

Marcy Winget
Fred Hutchinson Cancer Research Center
mwinget@fhcrc.org

Sean Kelly
NASA Jet Propulsion Laboratory
California Institute of Technology
sean.kelly@jpl.nasa.gov

Donald Johnsey
National Cancer Institute
National Institutes of Health
johnseyd@mail.nih.gov

Sudhir Srivastava
National Cancer Institute
National Institutes of Health
ss1a@nih.gov

Mark Thornquist
Fred Hutchinson Cancer Research Center
mthornqu@fhcrc.org

Abstract

Access to biospecimens is essential for enabling cancer biomarker discovery. The National Cancer Institute's (NCI) Early Detection Research Network (EDRN) comprises and integrates a large number of cancer research institutions into a network in order to establish a collaborative scientific environment to discover and validate disease markers. The diversity of both the institutions and the collaborative focus has created the need for establishing cross-disciplinary teams focused on integrating expertise in cancer research, computational and biostatistics, and computer science.

Given the collaborative design of the network, the EDRN needed an informatics infrastructure. The Fred Hutchinson Cancer Research Center, the National Cancer Institute, and NASA's Jet Propulsion Laboratory (JPL) teamed up to build an informatics infrastructure creating a collaborative, science-driven research environment despite the geographic and morphologic differences of the information systems that existed within the diverse network.

EDRN investigators identified the need to share biospecimen data captured across the country managed in disparate databases. As a result, the informatics team initiated an effort to create a virtual biospecimen database whereby scientists could search and retrieve details about specimens located at collaborating institutions. Each database, however, was locally implemented and integrated into collection processes and methods unique to each institution. This meant that efforts to integrate databases needed to be done in a manner that did not require redesign or re-implementation of existing systems.

1. Introduction

The Early Detection Research Network (EDRN) created and supported by the National Cancer Institute (NCI) is a five year, collaborative, multi-institutional scientific consortium [1]. The Network's goal is to identify, evaluate, and validate promising biomarkers to support the early detection of cancer. Access to biospecimens is essential for enabling the Network to obtain this goal. This paper is an update of an informatics infrastructure, ERNE, described previously by Crichton, et al. [2].

ERNE, the EDRN Resource Network Exchange, was developed to enable investigators to easily identify the availability of biospecimens and associated epidemiological information needed for their research. The system provides scientists access to biospecimen information regardless of where it is located across the country. ERNE's specific goal is to provide transparent access to existing specimen repositories providing EDRN a virtual knowledge environment despite the distributed nature of the collaboration. An overall informatics architecture and infrastructure was created for EDRN, plugging in databases that are managed locally by each institution.

The project focused on development of several key aspects including a common semantic architecture, a distributed informatics technology infrastructure that leveraged the semantic architecture, a dynamic portal, and a common study protocol for achieving compliance from each institution's Institutional Review Board (IRB). The project team took special care to minimize the impact of change and the requirements of high levels of informatics skills at each institution and to ensure that all data shared would be compliant with federal regulations.

Scientists use the system to search for and retrieve details about specimens located at collaborating sites. The ERNE system allows integration of local databases without altering collection processes and methods unique to each institution so that the burden of updating a central repository is not placed on anyone. Neither redesign nor re-implementation of existing systems is necessary.

The project team consists of the following institutions: the Data Management and Coordinating Center (DMCC) located at Fred Hutchinson Cancer Research Center serves as the project management and coordinating mechanism providing the central access point for the data management architecture; JPL provides the expertise and distributed software component infrastructure; and the NCI provides overall guidance. As of July 2003, nine EDRN sites are participating in this project to provide biospecimen repositories for querying. The ultimate goal is for all thirty EDRN institutions to be involved.

The software framework called the Object Oriented Data Technology (OODT) Framework [2] was provided by the Jet Propulsion Laboratory. OODT has been used by the National Aeronautics and Space Administration (NASA) for a wide variety of science disciplines including Planetary, Earth and Space Physics, as well as the informatics infrastructure for the ground data component of some of NASA's key missions. Most recently, OODT was used to provide the infrastructure for release of the planetary products from the 2001 Mars

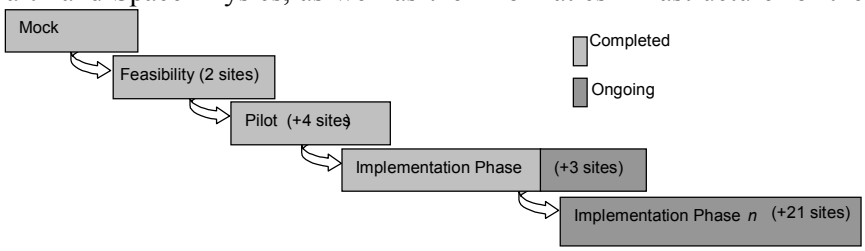

Figure 1. Multiphase plan and accomplishments to date for ERNE.

Odyssey mission. OODT provides a distributed component architecture that uses metadata as a means for integrating geographically distributed data resources.

The following were identified as key foci for this project: semantic architecture, informatics infrastructure, security and confidentiality, data model mapping, and a dynamic portal interface. Figure 1 illustrates the multi-phase plan and accomplishments to date for this project.

2. Semantic architecture

The underlying data models, both relationships and in many cases, terminology, of each system were locally defined making interoperability difficult. Many of the participating institutions had their own methods of representing data in their databases, presenting a real challenge for creating a virtual database. The EDRN developed a common ontology model for specimens and associated epidemiologic data that was useful in generating a set of common data elements (CDEs) for describing these data, their associated attributes, and relevant epidemiologic data. EDRN CDEs are data elements that have been agreed upon by the EDRN investigators as critical data that must be collected by all EDRN sites; they include data elements that describe study participants and specimens [3]. One of the key findings early on was that development of CDEs was essential for enabling interoperability across disparate databases. EDRN adopted the ISO/IEC 11179 (5e) standard [4]. This standard has provided a critical meta model framework for describing the CDEs in a consistent manner.

In order to develop the CDEs, the DMCC created working groups that combined multidisciplinary cancer research experts with computer scientists in an effort to establish the common language. Given that several sites had preexisting implementations using a different semantic architecture than EDRN, it became necessary to establish a mapping process that mapped the local data model at the institution to the CDEs. The DMCC continues to develop EDRN CDEs based on the ISO/IEC 11179 standard. This common model will continue to define a standard language for EDRN that will be used in all data sharing, data collection and informatics efforts.

3. Informatics infrastructure

In addition to creating a common model for describing biospecimens, the project team created an informatics software infrastructure, deployed via the Internet at all collaborating institutions in order to find and access information about specimens located in each institution's database. The system employed a metadata-based distributed framework as a synchronous communications infrastructure that tied databases together using the CDEs. Developing the common middleware allowed for data, normally tightly coupled to applications, to be de-coupled and integrated as a set of virtual repositories. In middleware, a request broker manages service requests from top tier client applications to server applications. Two server applications, the profile and product servers, provide search and retrieval functions and interface to catalogs and data repositories in the bottom tier of the architecture. This distributed framework makes it possible to query multiple institutional databases concurrently, compiling the results into a unified view of the available specimens [2].

Message-driven processing software (middleware) uses a request broker to handle service requests from clients to server applications. The message-driven paradigm addresses both interface as well as scalability issues since the number of component interconnections increases linearly as new components are added. The Object Oriented Data Technology (OODT) framework [5] is the foundation for the EDRN informatics infrastructure and provides the

messaging mechanism, product and profile servers, distributed server management, and plug-in capabilities for user tools. The EDRN ERNE middleware is configured as a single downloadable package and was installed at every site participating in the EDRN Informatics Project.

The software leverages Java's Remote Method Invocation (RMI) to support the distributed object implementation. This enables a common messaging layer within the system that all distributed servers use to communicate. The OODT distributed framework is designed in such a way as to support various distributed messaging implementations including Java RMI, Common Object Request Broker Architecture (CORBA), and Sun's new Peer-to-Peer implementation called JXTA [7]. These distributed messaging implementations provide services for distributed communication as well as object naming in order to locate distributed objects.

Product servers provide a common system interface to differing data repositories for data product access and retrieval. Each EDRN site downloads the software package and installs the product server component. Each product server runs a dynamically loaded Java object called a query handler that negotiates the interface between the EDRN enterprise environment and the local biospecimen repository. The query handler converts an EDRN query into a local query for the database. In general, this is a conversion to an SQL-compliant database (although any transformation is possible) that translates the query from EDRN CDEs to entity and attribute definitions defined by the local database model. Results from the query are mapped to the EDRN CDEs and then formatted with an agreed-upon representation. Any user application can request service from a product server through a standard HTTP or Java Application Programmer Interface (API), as well as C/C++ APIs.

Profile servers provide a common mechanism for describing distributed resources. The profile server manages profiles—sets of resource definitions [5]—about distributed data systems and their products. A profile is a metadata description of the resources known by a node in the distributed framework. These resources are interfaces, data products, or other profile servers available in the integrated enterprise. Profiles may be grouped and served by more than one profile server. The query component ties this architecture together by providing and managing the traversal of the integrated digraph node architecture. It also interprets profile definitions that provide mappings between data system nomenclature. The query component also provides the facility to manage concurrent queries across multiple servers to improve performance.

For this implementation, one profile server was used to reference each of the product servers that were located at distributed sites across the country. Figure 2 shows the deployment of the product servers at each institution, while Figure 3 provides an overview of the geographic distribution planned for spring 2003. Profile and product server instantiations are

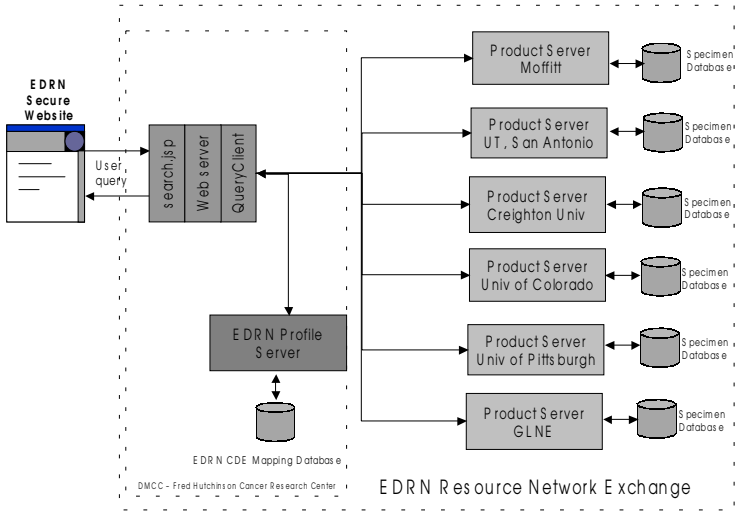

Figure 2. Software component deployment.

uniquely identified by name so they can be located within the distributed name server. These names are used as part of the metadata header encoded to identify the distributed EDRN services that can support queries for products.

The team chose XML since it provides a rich environment for defining and managing metadata.

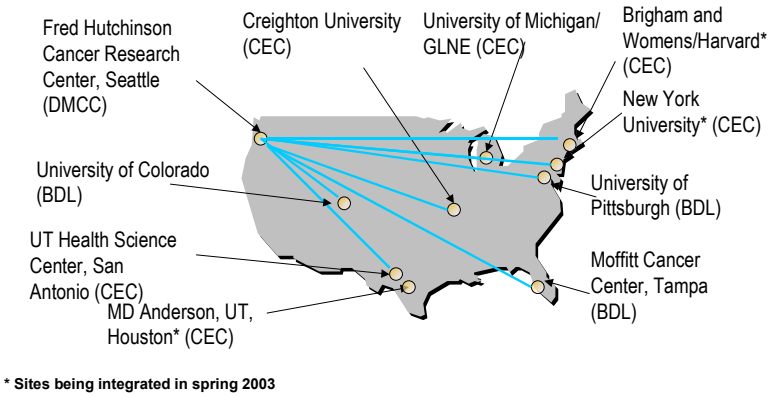

Figure 3. ERNE software installations.

In addition, XML serves as an interface specification on top of the distributed messaging layer between each of the nodes of the system. The query definition is implemented independent of any one database, functional, or programming language and is intended to provide an abstract view of both the query expression and the results. The query definition allows for each data system to be encapsulated. This allows various implementations, ranging from the use of relational and object database management systems to the use of flat file and home-grown databases for cataloging and storing data products to exchange information by plugging into a generic query definition.

One of the goals of this architecture is to provide a standard API that will allow for generic science analysis tools to be written that can plug into the architecture to retrieve and correlate data from multiple data sources. For this phase, the team developed a web-based interface using Java Server Pages (JSP). The interface serves as a client of the data architecture and allows researchers to query distributed databases from a single point.

The team discovered that architectural goals for space science and biomedical research were very similar, and in fact, the components developed for space science could be directly infused into the EDRN knowledge environment. By focusing on a framework for supporting basic system interoperability, the architecture was able to provide solutions that not only solve problems within a single discipline, but also support integration of cross-disciplinary databases.

4. Security and confidentiality

Specimen and associated participant data is highly sensitive information. Security and confidentiality were important considerations for this project, especially since the software transfers query results over the public Internet. Security requirements needed to be met for both the participating institutions and the federal Health Insurance Portability and Accountability Act (HIPAA) [6]. The system accommodated the local network configurations at various institutions in order to link them together via the Internet, protecting all network traffic using 128-bit strong encryption. Sites opened specific ports in local firewalls allowing traffic between product servers located at participating institutions and the DMCC. In addition, a template protocol was written to describe the project including information about each participating institution's technical environment, security infrastructure and database. This protocol was completed by each participating institution and approved by their Institutional Review Board (IRB). All Personal Health Information (PHI) identifiers have been removed from the data being shared to make them compliant with HIPAA regulations.

5. Data model mapping

Sharing of disparate representations of specimen banks could not have been possible without a high degree of communication. Sharing the EDRN CDEs that were used for the project with each participating site, as well as each participating site describing how their local data models mapped to the CDEs was a key component for the success of the project. The DMCC created an online CDE mapping tool, based on an MS Access tool developed by H. Lee Moffitt Cancer Center, to facilitate this process. This online tool resides on the EDRN secure website located at the DMCC. It allows each participating site to view the 39 EDRN CDEs that are being shared for this project as they are defined in the EDRN online metadata repository, and document how each of the data elements in their database maps or matches to a specific CDE. The tool also allows users to generate automatic emails to ease communication between groups. Each institution used the CDE mapping tool to document the transformations between their local database model and the EDRN CDE model.

This tool has allowed us to overcome several challenges in understanding the data contained in the local repositories and their corresponding data models. One such challenge was to communicate and document differences due to data collection prior to the establishment of the EDRN. For example, 'Gender' may be stored locally as 'M,' 'F,' and 'U' whereas the corresponding codes in the EDRN CDE data dictionary are '0,' '1,' and '9.' The mapping tool provides an explicit area to document the mapping and to indicate whether it corresponds to an exact match or not.

6. Dynamic portal interface

The team developed a unified portal using Java Server Pages (JSP) which connects the middleware to the distributed repositories. This portal allows scientists to specify search criteria using the projects specimen and epidemiological CDEs. The software translates the selections into an XML-based query definition, and two queries are sent. The first query verifies candidate sites that have a repository containing potential specimens that meet minimum criteria. The second query is then sent concurrently to the sites that may return results, as identified by the first query.

The portal is dynamic and will limit choices that appear based on specified user selections. When a user logs in they will see all sites and their operational status. If a site's server is down the user cannot select this site as part of the query. The portal also narrows choices based upon sites selected to show only the specimens that are available at those sites. The software summarizes results, showing the numbers of

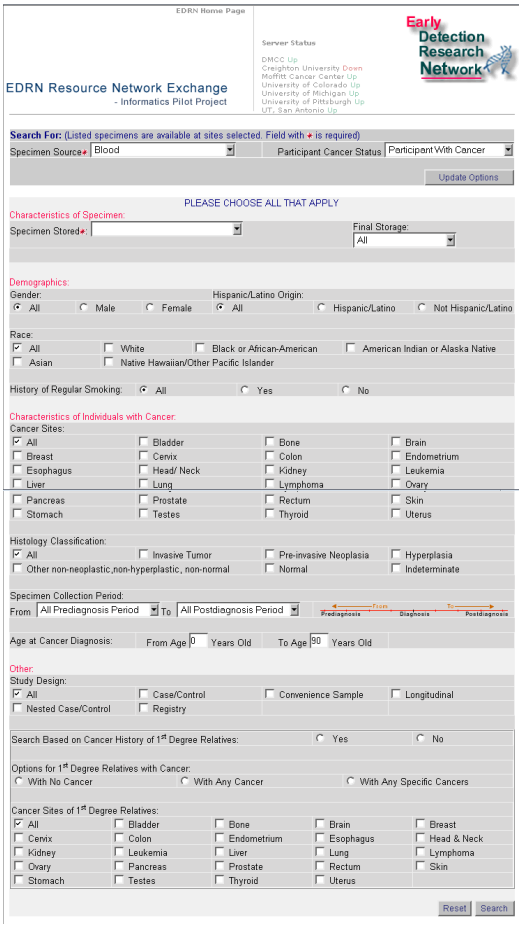

Figure 4. EDRN user interface

queried specimens available at each institution. Users can obtain more details regarding the specimen and characteristics of the donor by clicking on the "details" link adjacent to the site of interest. Data that identify the specimen donor are not available in any part of the results screens.

7. Conclusions

The achievements reflect the current progress of rolling out a national virtual specimen database to be used to enhance early cancer detection. We have achieved significant maturity in both the software infrastructure and in our team experience through this process and we envision that deployment of the architecture to support data access and sharing of similar data sets of specimens will occur far and wide. Developing methodologies for deploying software nationally presents challenges on several technical fronts. There are also policy and cultural challenges to overcome. As we move forward, we continue to change the paradigm by which data are delivered to scientists moving towards the goal of building a national knowledge infrastructure for cancer research. Future phases will target the creation of data profiles that include metadata descriptions of biospecimens, biomarkers including assay sensitivity and specificity, research protocols, and publications that ultimately will yield the biomarker knowledge environment.

Finally, our experience continues to show that the development of common metadata models is critical to development of a data architecture for biomedical research. Combining the data architecture with the Object Oriented Data Technology (OODT) technology architecture provides a platform for leading research institutions around the country to plug into enabling seamless access to specimens regardless of what institution manages them.

8. References

[1] S. Srivastava, B. Kramer, Early Detection Cancer Research Network: Lab Invest 80:1147-1148, 2000.

[2] D. Crichton, G. Downing, .H. Kincaid, S. Hughes, S. Srivastava, An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network, The 14th IEEE Conference on Computer Based Medical Systems, Institute of Electrical Engineers, July 26, 2001.

[3] M. Winget, J. Baron, M. Spitz, D. Brenner, D. Warzel, H. Kincaid, M. Thornquist, Z. Feng, International Journal of Medical Informatics, 70(1):41-48, 2003.

[4] ISO/IEC 11179 - Specification and Standardization of Data Elements, Parts 1-6, ISO/IEC specification, http://www.iso.ch/iso.

[5] D.J. Crichton, J.S. Hughes, J.J. Hyon, S.C. Kelly, Science Search and Retrieval using XML, The Second National Conference on Scientific and Technical Data, U.S. National Committee for CODATA, National Research Council, March 13-14, 2000, http://oodt.jpl.nasa.gov/doc/papers/codata/paper.pdf.

[6] Health Insurance Portability and Accountability Act. Standards for Privacy of Individually Identifiable Health Information - Rules and Regulations. Federal Register 65:250, 2000.

[7] L. Gong, Project JXTA: A Technology Overview. Sun Microsystems Incorporated, October 2002.

9. Acknowledgements

We wish to thank the following partner institutions for their involvement in this project: H. Lee Moffit Cancer Research Center, University of Texas at San Antonio, Creighton University, University of Colorado, University of Pittsburgh, Dartmouth Medical School, Brigham and Women's Hospital, University of Texas MD Anderson, and New York University.